Table of Contents

Including Monitoring traffic flows

General Conclusions

1. Traffic is not equally distributed to all monitoring servers if #felixServers is not divisible by #monServers (because of the grouping of Nfelix to 1 monServer).
2. Simulation shows very very slight increase in the data-flow latency compared to the scenario without monitoring (average increased ~3us , max peacks increased 50us).
3. HTL routers do present a bigger queueing effect as expected, although at this monitoring rate is minium as the link usage is < 50%

Simulated Scenario

Based on this meeting:
https://docs.google.com/presentation/d/1N5GnG82JvsASUJ6sl-gmuMBRjMWG6OWtybIjOMikb-M/edit#slide=id.g13

We agreed to add the full path of the monitoring traffic all the way to the monitoring servers in the HLT farm.

Monitoring flows (NFelix -> 1 MonServer

Topology Description (topologyGenerator)

NUMBER_OF_FELIX_SERVERS = 13 # this generates 1:1 connections with sw_rod, so NUMBER_OF_FELIX_SERVERS=numberOfSWRODServers
NUMBER_OF_MONITORING_SERVERS = 5
LINK_BW_40G_BITS_S = 40 * G # 40 Gbps
LINK_BW_1G_BITS_S = 1 * G # 1 Gbps
FELIX_FLOW_PRIORITY = 0
FELIX_GBT_ELINKS = 10 # #GBT e-links in each felix server. There will be one flow created per e-link (because there will be 1 thread, one connection per e-link)

# felix data-flow distributions (one per GBT)
FELIX_GBT_PERIOD_sec = ExponentialDistribution.new 1.0 / (100*K) # distribution period in seconds
#FELIX_GBT_PERIOD_sec = ExponentialDistribution.new 1.0 / (100) #TODO: this is just for testing quick simulations. Normal rate should be 100*k
FELIX_GBT_SIZE_bytes = NormalDistribution.new 4.0*K, 1.0*K # (in bytes)
FELIX_GBT_BUFFER_bytes = 1*M # (in bytes)
FELIX_GBT_TIME_OUT_sec = 2 # (in seconds)
FELIX_GBT_OUT_SIZE_bytes = TCP_MTU_bytes # (in bytes)

FELIX_GENERATION_PERIOD = FelixDistribution.new FELIX_GBT_PERIOD_sec,
FelixDistribution::FELIX_MODE_HIGH_THROUGHOUT,
FELIX_GBT_SIZE_bytes,
FELIX_GBT_BUFFER_bytes,
FELIX_GBT_TIME_OUT_sec,
FELIX_GBT_OUT_SIZE_bytes
FELIX_GENERATION_SIZE = ConstantDistribution.new TCP_MTU_bytes*8 #distribution size in bits

# monitoring flows (one per GBT)
FELIX_MONITORING_PRIORITY = 0

```
MONITORING_SIZE_bits = (TCP_MTU_bytes - 300)*8
TOTAL_MONITORING_PER_SERVER_bits = 0.3 * G
MONITORING_GENERATION_PERIOD = ExponentialDistribution.new 1.0 /
(TOTAL_MONITORING_PER_SERVER_bits / (MONITORING_SIZE_bits * FELIX_GBT_ELINKS))
MONITORING_GENERATION_SIZE = NormalDistribution.new MONITORING_SIZE_bits, 300*8
#distribution size in bits
```

Single simulation - Full Scenario

We fist started with a single simulation to validate results are as expected. We configured 0.3 Gb/s of monitoring traffic per felix server in the topologyGenerator. The rest of parameters as stated before for the full scenario: 13 felix servers (generating 32Gbps data-flow each, 10GBT), 5 monitoring servers, adding the HLT section for monitoring

Configuration (topologyGeneratior)

```
# monitoring flows (one per GBT)
FELIX_MONITORING_PRIORITY = 0
MONITORING_SIZE_bits = (TCP_MTU_bytes - 300)*8
TOTAL_MONITORING_PER_SERVER_bits = 0.3 * G
MONITORING_GENERATION_PERIOD = ExponentialDistribution.new 1.0 /
(TOTAL_MONITORING_PER_SERVER_bits / (MONITORING_SIZE_bits * FELIX_GBT_ELINKS))
MONITORING_GENERATION_SIZE = NormalDistribution.new MONITORING_SIZE_bits, 300*8
#distribution size in bits
```

Simulation Results

backup: /afs/cern.ch/work/m/mbonaven/public/SimuResults/PhaseI/LArSlice_1_to_1/LAr_withMonitoring
git commit: 08133de2

NOTE: In some figures we only include plots of some of the nodes to make them more readable. For example, instead of plotting the 13 swrods, we only include the plot of the first 2 and the last 2. The rest of the servers were validated but not included here.

Throughput

SWRODs

With 10GBT links (each generating at 400MB/s=100KHz * 4KB) each felix server generates 4GB/s. It is expected for each SW_ROD to receive this data as no congestion is expected.

monitoring servers

The plot shows expected throughput at the monitoring servers:

- monitoring servers 0, 1 and 2 received ~113MB/s=3*37.5MB/s.
- monitoring servers 3 and 4 received ~75MB/s=2*37.5MB/s.

With 13 felix servers and 5 monitoring servers, monitoring servers 0, 1 and 2 will receive traffic from 3 felix servers each while monitoring servers 3 and 4 will only receive traffic from 2 felix each. Each felix servers is configured to generate 0.3Gb/s=37.5MB/s of monitoring.

Latency

SWRODs

The plot shows very slight increase in the data-flow latency compared to the scenario without monitoring (average increased ~3us , max peacks increased 50us). Without monitoring the network latency was in average 139us with max peaks oof 250us. With monitoring the network latency increased to an average of 142us with max peaks of 300us

Monitoring servers

Plot shows that the mean average latency for the monitoring traffic is ~80us. Equally for all monitoring servers, maybe servers 3 and 4 with slightly less latency as they have one less felix server sending data.

The minimum theoretical latency is ~0.1us. This suggests that there is a queing effect

Link Usage

In this scenario we add the plots for the extra felix servers (13 in total, only 4 ploted), the felix routers(2 routers), and the HLT routers (2 routers)

Felix servers

As expected both links are equally used in average (~2.01GB/s). This corresponds to 4GB/s of data flow + 0.037 GB/s of monitoring, which is equally dividen the the 2 outlinks (bonded link). 4.037 GBps / 2 = 2.018 GBps

Felix switches/routers

Compared to the scenario with monitoring the switches present the same link usage down to the SWRODs (as expected). For the link down to the cores (port0), they now show a usage of 243MB/s each of the 2 swithes. 243Mb/s corresponds to the monitoring traffic of all 13 felix servers (37.5MB/s * 13 = 487.5MB/s) which is router half to one swithc and half to the other.

For the felix routers they only have a single port each, which sends all the 243MB/s down to the HLT routers.

HLT routers

HLT routers receive 243MB/s each from the felix network corresponding to the monitoring traffic. This traffic is routed to the 5 monitoring servers.

ports 0, 1 and 2 are connected to monitoring servers 0, 1 and 2 which receive traffic from 3 felix servers. ports 3 and 4 route packets from 2 felix servers.

Queue sizes

Note on the queue plots: the figures plot the **MAXIMUM** queue size in a given sampling period (in this case samplingPeriod=0.01s). This is because we are interested in the queue required to achieve no discards. In the legends of the figures, the **TIME _ AVERAGE** is shown: this is the queue size average taking time into consideration => queueSize{i} / totalTimeWithSize{i}. See SamplerLogger and TimeAvg definition for more details.

Felix servers

The output queues in the felix server NICs are very very slighty increased compared to the scenario without traffic (before max peak of 3.5MB, now 4MB. Before avgLenght of 270kB, now 290kB).

The big difference between the maxium usage and the avgUsage denotes that the queues increase with bursts (comming from the flushing of the buffers) and quickly emptied.

Felix switches/routers

lar_switches show a minimal queing (timeAvg: 6.6B, maxQueue: 12KB) which is caused by agregation into a single link of the monitoring traffic incomming from the different felix servers. Although there is a big aggregation of links (13:1), the monitoring traffic is a very small percentaje of the link capacity.

On the felix_core routers there is an even smaller queueing effect (timeAvg: 3.5B, maxQueue: 2500 B) whiich corresponds to ~1-2 packets (monitoring packet size avg: 1200B).

HLT routers

HTL routers do present a bigger queueing effect as expected, although at this monitoring rate is minium as the link usage is < 50%
There is a maximum at the very beggining of the simulation of 45 KB (why higher at the beggining??) and then a maximum of ~25KB. The time average is ~0.9KB or ~0.4KB, which suggests buffers are empied very quickly.

Performance

Number of generated packets: ??? M ==>

Simulation execution:

- Initialization TOTAL time: 256810 (ms) [in 2 felix no-monitoring scenario 40398 (ms)]
- Simulation time (not including init): 4271652 (ms) [in 2 felix no-monitoring scenario 93930 (ms)]
- **TOTAL execution time: 4530020 (ms) [in 2 felix no-monitoring scenario 134649 (ms)]**

=> ~0.05ms of execution per simulated packet (not including init time) **[in basic scenario 0.035 (ms)]**

Compared to the basic scenario:

- initialization time : ???
- Simulation time per packet : ??

-- MatiasAlejandroBonaventura - 2016-12-08

This topic: Main > PhaseISimulation_LArSliceMonitoringTraffic
Topic revision: r2 - 2016-12-08 - MatiasAlejandroBonaventura